PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) **International Patent Classification** 6 : **G10L 5/06**

A2

(11) **International Publication Number:** **WO 99/31654**

(43) **International Publication Date:** 24 June 1999 (24.06.99)

(21) **International Application Number:** PCT/IB98/01990

(22) **International Filing Date:** 11 December 1998 (11.12.98)

(30) **Priority Data:**

197 55 191.2	12 December 1997 (12.12.97)	DE
98203725.1	6 November 1998 (06.11.98)	EP

(81) **Designated States:** JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) **Applicant** *(for all designated States except US)*: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL–5621 BA Eindhoven (NL).

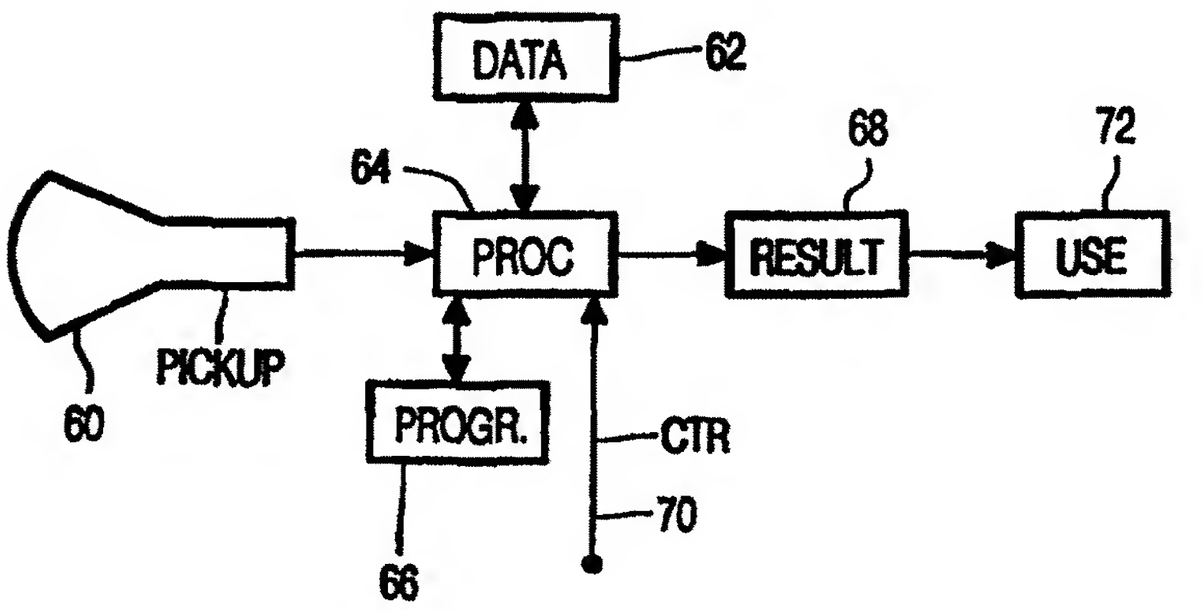
(71) **Applicant** *(for SE only)*: PHILIPS AB [SE/SE]; Kottbygatan 7, Kista, S–164 85 Stockholm (SE).

(71) **Applicant** *(for DE only)*: PHILIPS PATENTVERWALTUNG GMBH [DE/DE]; Röntgenstrasse 24, D–22335 Hamburg (DE).

(72) **Inventor; and**
(75) **Inventor/Applicant** *(for US only)*: BEYERLEIN, Peter [DE/DE]; Prof. Holstlaan 6, NL–5656 AA Eindhoven (DE).

(74) **Agent:** GÖSSMANN, Klemens; Internationaal Octrooibureau B.V., P.O. Box 220, NL–5600 AE Eindhoven (NL).

(54) **Title:** METHOD OF DETERMINING MODEL–SPECIFIC FACTORS FOR PATTERN RECOGNITION, IN PARTICULAR FOR SPEECH PATTERNS



(57) **Abstract**

A method for recognizing a pattern that comprises a set of physical stimuli, said method comprising the steps of: providing a set of training observations and through applying a plurality of association models ascertaining various measuring values $p_j(k \acute{Y} x)$, j=1... M, that each pertain to assigning a particular training observation to one or more associated pattern classes; setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and joining thereto a normalization quantity to produce a compound association distribution; optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution; recognizing target observations representing a target pattern with the help of said compound distribution.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Method of determining model-specific factors for pattern recognition, in particular for speech patterns.

BACKGROUND OF THE INVENTION

The invention relates to a method for recognizing a pattern that comprises a set of physical stimuli, said method comprising the steps of:

- providing a set of training observations and through applying a plurality of association models ascertaining various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning a particular training observation to one or more associated pattern classes;
- setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and joining thereto a normalization quantity to produce a compound association distribution.

The invention has been conceived for speech recognition, but is likewise applicable to other recognition processes, such as for speech understanding, speech translation, as well as for recognizing handwriting, faces, scene recognition, and other environments. The association models may be probability models that give probability distributions for assigning patterns to classes. Other models can be based on fuzzy logic, or similarity measures, such as distances measured between target and class. Known technology has used different such models in a combined recognition attack, but the influences lent to the various cooperating models were determined in a haphazard manner. This meant that only few and/or only elementary models were feasible.

The present inventor has recognized that the unification of Maximum-Entropy and Discriminative Training principles would in case of combination of more than one model in principle be able to attain superior results as compared with earlier heuristic methods. Also, a straightforward data processing procedure should provide a cheap and fast road to those results.

In consequence, amongst other things, it is an object of the invention to evaluate a log-linear combination of various 'sub'models $p_j(k \mid X)$ whilst executing parameter evaluation through discriminative training. Now, according to one of its aspects, the invention attains the object by recognizing a pattern that comprises a set of physical stimuli, said method comprising the steps of:

- providing a set of training observations and through applying a plurality of association

models ascertaining various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning a particular training observation to one or more associated pattern classes;

- setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and joining thereto a normalization quantity to produce a compound association distribution;

- optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution;

- recognizing target observations representing a target pattern with the help of said compound distribution. Inter alia, such procedure allows to combine any number of models into a single maximum-entropy distribution. Furthermore, it allows an optimized interaction of models that may vary widely in character and representation.

The invention also relates to a method for modelling an association distribution according to the invention. This provides an excellent tool for subsequent users of the compound distribution for recognizing appropriate patterns.

The invention also relates to a method for recognizing patterns using a compound distribution produced by the invention. This method has users benefitting to a great deal by applying the tool realized by the invention.

The invention relates to a system that is arranged for practising a method according to the invention. Further aspects are recited in dependent Claims.

BRIEF DESCRIPTION OF THE DRAWING

These and other aspects and advantages of the invention will be discussed more in detail with reference to the detailed disclosure of preferred embodiments hereinafter, and in particular with reference to the appended Figures that show:

Fig. 1, an overall flow chart of the method;

Fig. 2, a comprehensive system for practising the invention;

Figures 3-21 give various equations B1-B20 used with the automatic method according to the invention.

DETAILED DISCLOSURE OF PREFERRED EMBODIMENTS

The invention being based on a balanced application of mathematics on the handling and accommodating of physical quantities that may be of very diverse character, much of the disclosure is based on advanced mathematics. However, both the starting point and the eventual outcome have permanently physical aspects and relevance. The speech

recognition may be used to control various types of machinery. Scene analysis may guide unmanned vehicles. Picture recognition may be used for gate control. Various other applications are evident per se. The expressions hereinafter are numbered in sequence, and will be referred to in the text by these numbers.

The invention determines model-specific factors in order to combine and optimize several different models into a single pattern recognition process, notably for speech recognition.

The statistical speech recognition method utilizes Bayes' decision theory in order to form an identification mechanism with a minimum error rate. In conformity with this theory, the decision rule is such, that an observation x must be assigned to the class k (x ϵ k for brevity), when for a given a posteriori or "real" probability distribution $\pi(k|x)$ it holds that:

$$\forall k' = 1,\ldots,K;\ k' \neq k : \log \frac{\pi(k|x)}{\pi(k'|x)} \geq 0 \Rightarrow x \in k \qquad (1)$$

In literature, the term $\log(\pi(k|x)/\pi(k'|x))$ is called the discriminant function. Hereinafter, this term will be noted g(x,k,k') will be used for brevity. When the decision rule (1) is used for recognizing complete sentences, observed expressions $x_1^T = (x^1,\ldots,x^T)$, that have a temporal length T, will be classified as spoken word sequences $w_1^S = (w^1,\ldots,w^S)$ of length S. The a posteriori distribution $\pi(w_1^S|x_1^T)$ is however unknown since it describes the complicated natural speech communication process of humans. Consequently, it must be approximated by a distribution $p(w_1^S|x_1^T)$. Thus far, the acoustic-phonetic and grammatical modelling of speech in the form of parametric probability distributions have attained the best results. The form of the distribution $p(w_1^S|x_1^T)$ is then predetermined; the unknown parameters of the distribution are estimated on the basis of training data. The distribution $p(w_1^S|x_1^T)$ so acquired is subsequently inserted into Bayes' decision rule. The expression x_1^T is then assigned to the word sequence w_1^S for which:

$$\forall w'^{s'}_1 \neq w^s_1 : \log\frac{p(w^s_1|x^T_1)}{p(w'^{s'}_1|x^T_1)} > 0 \qquad (2)$$

Conversion of the discriminant function

$$g(s^T_1, w^s_1, w'^{s'}_1) = \log\frac{p(w^s_1|x^T_1)}{p(w'^{s'}_1|x^T_1)} = \log\frac{p(w^s_1)p(x^T_1|w^s_1)}{p(w'^{s'}_1)p(x^T_1|w'^{s'}_1)}, \qquad (3)$$

allows to separate the grammatical model $p(w^s_1)$ from the acoustic-phonetic model $p(x^T_1|w^s_1)$ in a natural way. The grammatical model $p(w^s_1)$ then describes the probability of occurrence of the word sequence w^s_1 per se, and the acoustic-phonetic model $p(x^T_1|w^s_1)$ evaluates the probability of occurrence of the acoustic signal x^T_1 during the uttering of the word sequence w^s_1. Both models can then be estimated separately, so that an optimum use can be made of the relatively limited amount of training data. The decision rule (3) could be less than optimum due to a deviation of the distribution p from the unknown distribution π, even though the estimation of the distribution p was optimum. This fact motivates the use of so-called discriminative methods. Discriminative methods optimize the distribution p directly in respect of the error rate of the decision rule as measured empirically on training data. The simplest example of such discriminative optimization is the use of the so-called language model factor λ. The equation (3) is then modified as follows:

$$g(x^T_1, w^s_1, w'^{s'}_1) = \log\frac{p(w^s_1)^\lambda p(x^T_1|w^s_1)}{p(w'^{s'}_1)^\lambda p(x^T_1|w'^{s'}_1)} \qquad (4)$$

Experiments show that the error rate incurred by the decision rule (4) decreases when choosing $\lambda > 1$. The reason for this deviation from theory, wherein $\lambda = 1$, lies in the incomplete or incorrect modelling of the probability of the compound event (w_1^s, x_1^T). The latter is inevitable, since the knowledge of the process generating the event (w_1, x_1^T) is incomplete.

Many acoustic-phonetic and grammatical language models have been analyzed thus far. The object of these analyses was to find the "best" model for the relevant recognition task out of the set of known or given models. All models determined in this manner are however imperfect representations of the real probability distribution, so that when these models are used for pattern recognition, such as speech recognition, incorrect recognitions occur as incorrect assignment to classes.

It is an object of the invention to provide a modelling, notably for speech, which approximates the real probability distribution more closely and which nevertheless can be carried out while applying only little processing effort, and in particular to allow easy integration of a higher number of known or given models into a single classifier mechanism.

SUMMARY OF THE INVENTION

The novel aspect of the approach is that it does not attempt to integrate known speech properties into a single acoustic-phonetic distribution model and into a single grammatical distribution model which would involve complex and difficult training. The various acoustic-phonetic and grammatical properties are now modeled separately and trained in the form of various distributions $p_j(w_1^S | x_1^T)$, $j=1...M$, followed by integration into a compound distribution

$$p_{\{\Lambda\}}^{\pi}(w_1^s | x_1^T) = C(\Lambda).\prod_{j=1}^{M} p_j(w_1^s | x_1^T)^{\lambda_j} = \exp\left\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log p_j(w_1^s | x_1^T)\right\} \quad (5)$$

The effect of the model p_j on the distribution $p^{\pi}_{\{\Lambda\}}$ is determined by the associated coefficient λ_j.

The factor $C(\Lambda)$ ensures that the normalization condition for probabilities is satisfied. The free coefficients $\Lambda = (\lambda_1,\ldots,\lambda_M)^{tr}$ are adjusted so that the error rate of the resultant discriminant function

$$g(x_1^T, w_1^s, w'^{s'}_1) = \log \frac{\prod_{j=1}^{M} p_j(w_1^s | x_1^T)^{\lambda_j}}{\prod_{j=1}^{M} p_j(w'^{s'}_1 | x_1^T)^{\lambda_j}} \qquad (6)$$

is as low as possible. There are various possibilities for implementing of this basic idea, several of which will be described in detail hereinafter.

First of all, various terms used herein will be defined. Each word sequence w_1^s forms a class k; the sequence length S may vary from one class to another. A speech utterance x_1^T is considered as an observation x; its length T may then differ from one observation to another.

Training data is denoted by the references (x_n, k), with $n = 1,\ldots,N$; $k = 0,\ldots,K$. Herein N is the number of acoustic training observations x_n, and k_n is the correct class associated with the observation x_n. Further, $k \neq k_n$ are the various incorrect rival classes that compete with respect to k_n.

The classification of the observation x into the class k in conformity with Bayes' decision rule (1) will be considered. The observation x is an acoustic realization of the class k. In the case of speech recognition, each class k symbolizes a sequence of words. However, the method can be applied more generally.

Because the class k_n produced by the training observation x_n is known, an ideal empirical distribution $\hat{\pi}(k|x)$ can be constructed on the basis of the training data (x_n, k); n=1... N; k=0...K. This distribution should be such that the decision rule derived therefrom has a minimum error rate when applied to the training data. In the case of classification of complete word sequences k, a classification error through selecting an erroneous word sequence $k \neq k_n$, may lead to several word errors. The number of word errors between the incorrect class k and the correct class k_n is called the Levenshtein distance

$E(k,k_n)$. The decision rule formed from $E(k,k_n)$ has a minimum word error rate when a monotony property is satisfied.

The ideal empirical distribution $\hat{\pi}$ is a function of the empirical error value $E(k,k_n)$ which is given only for the training data, but is not defined with respect to unknown test data, because the correct class assignment is not given in that case. Therefore, on the basis of this distribution there is sought a distribution

$$p^{\pi}\{\Lambda\}(k|x) = \frac{\exp\left\{\sum_{j=1}^{M} \lambda_j \log p_j(k|x)\right\}}{\sum_{k'=1}^{K} \exp\left\{\sum_{j=1}^{M} \lambda_j \log p_j(k'|x)\right\}} \quad (7)$$

which is defined over arbitrary, independent test data and has an as low as possible empirical error rate on the training data. If the M predetermined distribution models $p_1(k|x),\ldots,p_M(k|x)$, are defined on arbitrary test data, the foregoing also holds for the distribution $p^{\pi}_{\{\Lambda\}}(k|x)$. When the freely selectable coefficients $\Lambda = (\lambda 1,\ldots, \lambda_M)^{tr}$ are determined in such a manner that $p^{\pi}_{\{\Lambda\}}(k|x)$ on the training data has a minimum error rate, and if the training data is representative, $p^{\pi}_{\{\Lambda\}}(k|x)$ should yield an optimum decision rule also on independent test data.

The GPD method as well as the least mean square method optimize a criterion which approximates the mean error rate of the classifier. In comparison with the GPD method, the least mean square method offers the advantage that it yields a closed solution for the optimum coefficient Λ.

The least mean square method will first be considered. Because the discriminant function (1) determines the quality of the classifier, the coefficients Λ should minimize the root mean square deviation B14 of the discriminant functions of the distributions $p^{\pi}_{\{\Lambda\}}(k|x)$ from the empirical error rate $E(k,k_n)$. The summing over r then includes all rival classes in the criterion. Minimizing D (Λ) leads to a closed solution for the optimum coefficient vector $\Lambda = Q^{-1}P$ (9), further detailed by B15 and B16.

Herein, Q is the autocorrelation matrix of the discriminant functions of the predetermined distribution models. The vector P expresses the relationship between the discriminant functions of the predetermined models and the discriminant function of the distribution $\hat{\pi}$.

The word error rate E(k, k_n) of the hypotheses k is thus linearly taken up in the coefficients $\lambda_1,\ldots,\lambda_M$. Conversely, the discrimination capacity of the distribution model p_i is linearly included in the coefficients $\lambda_1,\ldots,\lambda_M$ for determining the coefficients directly via the discriminant function $\log\frac{p_i(k|x_n)}{p_i(k_n|x_n)}$.

Alternatively, these coefficients can be determined by using the GPD method. With this method, the smoothed empirical error rate E(Λ):

$$E(\Lambda) = \frac{1}{N}\sum_{n=1}^{N} l(x_n,k_n,\Lambda) \tag{12}$$

$$l(x_n,k_{n0},\Lambda) = \left\{1 + A\left[\frac{1}{K}\sum_{r=1}^{K}\exp\left\{-\eta\log\frac{p^{\pi}_{\{\Lambda\}}(k_n|x_n)}{p^{\pi}_{\{\Lambda\}}(k|x_n)}\right\}\right]^{-\frac{B}{\eta}}\right\}^{-1} \tag{13}$$

can be directly minimized for the training data. The left hand expression is then a smoothed measure for the error classification risk of the observation x_n. The values $A>0$, $B>0$, $\eta>0$ determine the type of smoothing of the error classification risk and should be suitably predetermined. When E(λ) is minimized in respect of the coefficient λ of the log linear combination, the following iteration equation with the step width M is obtained for the coefficients λ_j, wherein j = 1,...,M

$\lambda_j^{(0)} = 1$ (11), and furthermore according to B13 and B14, and

$\Lambda^{(I)} = (\lambda_1^{(I)},\ldots,\lambda_M^{(I)})^{tr}; j = 1,\ldots,M$

It is to be noted that the coefficient vector Λ is included in the criterion $E(\Lambda)$ by way of the discriminant function

$$\log \frac{p^{\pi}_{\{\Lambda\}}(k_n|x_n)}{p^{\pi}_{\{\Lambda\}}(k|x_n)} \tag{12}$$

If $E(\Lambda)$ decreases, the discriminant function (12) should increase on average because of (9) and (10). This results in a further improved decision rule, see (1).

In the above, the aim has been to integrate all available knowledge sources into a single pattern recognition system. Two principles are united. The first is the maximum-entropy principle. This works by introducing as few assumptions as possible, so that uncertainty is maximized. Thus, exponential distributions must be used. In this manner the structure of the sources combination is defined. The second principle is discriminative training, to determine the weighting factors assigned to the various knowledge sources, and the associated models. Through optimizing the parameters, the errors are minimized. For speech, models may be semantic, syntactic, acoustic, and others.

The approach is the log-linear combining of various submodels and the estimating of parameters through discriminative training. In this manner, the adding of a submodel may improve the recognition score. If not, the model in question may be discarded. A submodel can however never decrease the recognition accuracy. In this manner, all available submodels may be combined to yield optimum results. Another application of the invention is to adapt an existing model combination to a new recognition environment.

The theoretical approach of the procedure includes various aspects:

- parabolic smoothing of the empirical error rate
- simplifying the theory of "minimum error rate training"
- providing a closed form solution that needs no iteration sequence.

The invention furthermore provides extra facilities:

- estimating an optimum language model factor
- applying a log-linear Hidden Markov Model
- closed form equations for optimum model combination
- closed form equations for discriminative training of class-specific probability distributions.

Now for the classification task specified in (1), the true or posterior

distribution $\pi(k \mid x)$ is unknown but approximated by a model distribution ($p(k \mid x)$. The two distribution differ, because of incorrect modelling assumptions and because of insufficient data. An example is the language model factor λ used in equation B1.

The formal definition combines various submodels $p_j(k \mid x)$, $j=1...M$ into a log-linear posterior distribution $p_{\{\Lambda\}}(k \mid x) = \exp\{..\}$ as given in (5). Next to the log-linear combination of the various submodes, the term $\log C(\Lambda)$ allows normalization to attain a formal probability distribution. The resulting discriminant function is

$$\log \frac{p_{\{\Lambda\}}(k|x)}{p_{\{\Lambda\}}(k'|x)} = \sum_j \lambda_j \log \frac{p_j(k|x)}{p_j(k'|x)} \qquad \text{(B2)}$$

The error rate is minimized and Λ optimized. Optimizing on the sentence level is as follows:

- Class k: word sequence
- Observation x: spoken utterance (e.g. sentence)
- N training samples x_n, giving the correct sentence
- For each sample x_n
  - k_n: correct class as spoken
  - $k \neq k_n$: rival classes, which may be all possible sentences, or for example, a reasonable subset thereof.
- Similarity of classes: $E(k_n,k)$
  - E: suitable function of Levenshthein-Distance, or a similarly suitable measure that is monotonous.
- Number of words in wordsequence k_n: L_n.

Now, equation B3 gives an objective function, the empirical error rate. Herein, the left hand side of the equation introduces the most probable class that bases on the number of erroneous deviations between classes k and k_n.

The parameters Λ may be estimated by:

- a minimum error rate training through Generalized Probabilistic Descent, which yields an iterative solution.
- a modification thereof combines with parabolic smoothing, which yields a closed form solution.
- a third method bases on least squares, which again yields a closed form solution.

For the GPD method, the smoothed empirical error rate minimizing is

based on expression B4. The smoothed misclassification risk is given by equation B5, and the average rivalry by equation B6.

The smoothed empirical error ate is minimized through B7. Herein, l is a loss function that for straightforward calculations must be differentiatable. Rivalry is given by equation B8, wherein E indicates the number of errors. Average rivalry is given through the summing in equation B9. A smoothed misclassification risk is expressed by equation B10 that behaves like a sigmoid function. For $R_n=-\infty$,l becomes zero, for $R_n=+\infty$, the limiting value is l=1. Herein A, B are scaling constants greater than zero. Differentiating to Λ yields expression B11, wherein the vector $\Lambda^{(l)}$ is given by expression B12 and the eventual outcome by expression B13.

The invention also provides a closed form solution for finding the discriminative model combination DMC. The solution is to minimize the distance between on the one hand the discriminant function and on the other hand the ideal discriminant function $E(k_n,k)$ in a least squares method. The basic expression is given by equation B14. Herein, $\Lambda=Q^{-1}P$, wherein Q is a matrix with elements $Q_{i,j}$ given by equation B15. Furthermore, P is a vector with elements P_i given by equation B16. Now, the empirical error rate has been given earlier in equation B3. For calculatory reasons this is approximated by a smoothed empirical error rate as expressed by equation B17. Herein, an indication is given on the number of errors between k and k_n through using a sigmoid function S or a similarly useful function. A useful form is $S(x)=\{(x+B)/(A+B)\}^2$, wherein $-B<x<A$ and $-B<0<A$. for higher values of x, S=1, and for lower values S=0. This parabola has proved to be useful. Various other second degree curves have been found useful. The relevant rivals must now lie in the central and parabolically curved interval of S. Now, finally, a normalization constraint is added for Λ according to expression B18.

The second criterion is solved according to a matrix equation (α, $\lambda^{tr})^{tr}=Q'^{-1}P'$, wherein an additional row and column have been supplemented to matrix Q' for normalization reasons, according to $Q'_{0,0}=0$; $Q'_{0,j}=1$, $Q'_{i,0}=1/2(A+B)^2$. The general element of correlation matrix Q' has been given in equation B19. Note that the closed solution is rendered possible through the smoothed step function s. Furthermore, vector P' likewise gets a normalizing element $P'_0=1$, whereas its general element is given by expression B20.

Experiments have been done with various M-gram language models, such as bigram, trigram, fourgram or tetragram models, various acoustic models, such as word-

internal-triphone, cross-word-trigram and pentaphone models. Generally, the automatic DMC procedure performs equally well as the results produced by non-automatic fine tuning using the same set of submodels. However, the addition of extra submodels according to the automatic procedure of the invention allowed to decrease the number of errors by about 8%. This is considered a significant step forward in the refined art of speech recognition. It is expected that the invention could provide similarly excellent results for recognizing other types of patterns, such as signatures, handwriting scene analysis, and the like, given the availability of appropriate sub-models. Other submodels used for broadcast recognition included **mllr** adaptation, unigram, distance-1 bigram, wherein an intermediate element is considered as don't care, pentaphones and wsj-models. In this environment, raising the number of submodels in the automatic procedure of the invention also lowered the numbers of errors by a significant amount of 8-13%.

Figure 1 shows an overall flow chart of a method according to the invention. In block 20 the training is started on a set of training data or patterns that is provided in block 22. The start as far as necessary claims required software and hardware facilities; in particular, the various submodels and the identity of the various patterns is also provided. For simplicity, the number of submodels has been limited to 2, but the number may be higher. In parallel blocks 24 and 26, the scores are determined for the individual submodels. In block 28 the log-lin combination of the various submodels is executed and normalized. In block 30 the machine optimizing of vector Λ in view of the lowest attainable error rate is executed. Note that vector Λ may have one or more zero-valued components to signal that the associated submodel or -models would bring about no improvement at all.

Next, the vector Λ and the various applicable submodels will be used for recognizing target data, as shown in the right half of the Figure. The training at left, and the usage at right may be executed remote from each other both in time and in space; for example a person could have a machine trained to that person's voice at a provider's premises. This might require extra data processing facilities. Later, the machine so trained may be used in a household or automobile environment, or other. Thus, blocks 40-46 have corresponding blocks at left.

In block 48 the scorings from the various submodels are log-lin combined, using the various components of vector Λ that had been found in the training. Finally, in block 50 the target data are classified using the results from block 50. In block 52, the procedure is stopped when ready.

Figure 2 shows a comprehensive system for practising the invention. The

necessary facilities may be mapped on standard hardware, or on a special purpose machine. Item 60 is an appropriate pickup, such as a voice recorder, a two-dimensional optical scanner, together with A/D facilities and quality enhancing preprocessing if necessary. Block 64 represents the processing that applies programs from program memory 66 on data that may arrive from pickup 60, or from data storage 62, where they may have been stored permanently or transiently, after forwarding from pickup 60. Line 70 may receive user control signals, such as start/stop, and possibly training-supportive signals, such as for definitively disabling a non-contributory submodel.

Block 68 renders the recognition result usable, such as by tabulating, printing, addressing a dialog structure for retrieving a suitable speech answer, or selecting a suitable output control signal. Block 72 symbolizes the use of the recognized speech, such as outputting a speech riposte, opening a gate for a recognized person, selecting a path in a sorting machine, and the like.

$$\log \frac{\pi(x|k) \cdot \pi(k)}{\pi(x|k') \cdot \pi(k')} \quad \rightarrow \quad \log \frac{p(x|k) \cdot p(k)^{\lambda}}{p(x|k') \cdot p(k')^{\lambda}}$$

B 1

$$\frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \mathsf{E}\left(\arg\max_{k} \left(\log \frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)} \right), k_n \right) =$$

$$\frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathsf{E}(k_n, k) \cdot \delta\left(k \ , \ \arg\max_{k'} \left(\log \frac{p_{\wedge}(k'|x_n)}{p_{\wedge}(k_n|x_n)} \right) \right)$$

B 3

$$L(\wedge) \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, \wedge)$$

B 4

$$\ell(x_n, \Lambda) = \frac{1}{1 + AR_n(\Lambda)^{-B}}$$

B 5

$$R_n(\Lambda) = \left(\frac{1}{K-1} \sum_{k \neq k_n} \left(e^{\mathsf{E}(k_n,k) \cdot \sum_{j=1}^{M} \lambda_j \log \frac{p_j(k|x_n)}{p_j(k_n|x_n)}} \right)^{\eta} \right)^{\frac{1}{\eta}}$$

B 6

$$L(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, \Lambda)$$

B 7

$$y_{nk}(\Lambda) = \left(\frac{p_{\Lambda}(k|x_n)}{p_{\Lambda}(k_n|x_n)} \right)^{\mathsf{E}(k_n,k)}$$

B 8

$$R_n(\Lambda) = \left(\frac{1}{K-1} \sum_{k \neq k_n} [y_{nk}(\Lambda)]^{\eta} \right)^{\frac{1}{\eta}} \quad \text{B 9}$$

$$\ell(x_n, \Lambda) = \frac{1}{1 + AR_n(\Lambda)^{-B}} \quad \text{B 10}$$

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} - \varepsilon \sum_{n=1}^{N} \ell(x_n, \Lambda^{(I)}) \left(1 - \ell(x_n, \Lambda^{(I)})\right) \times$$

$$\frac{\sum_{k \neq k_n} \mathsf{E}(k_n, k) \log \left(\frac{p_j(k|x_n)}{p_j(k_n|x_n)} \right) [y_{nk}(\Lambda^{(I)})]^{\eta}}{\sum_{k \neq k_n} [y_{nk}(\Lambda^{(I)})]^{\eta}} \quad \text{B 11}$$

$$\Lambda^{(I)} = (\lambda_1^{(I)}, \ldots, \lambda_M^{(I)})^{tr} \quad \text{B 12}$$

$$y_{nk}(\Lambda) = \left(\frac{p_{\Lambda}(k|x_n)}{p_{\Lambda}(k_n|x_n)} \right)^{\mathsf{E}(k_n,k)} \quad \text{B 13}$$

$$\frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left(\log\frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)}-\mathsf{E}(k_n,k)\right)^2 =$$

$$\frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left(\sum_j \lambda_j \log\frac{p_j(k|x_n)}{p_j(k_n|x_n)}-\mathsf{E}(k_n,k)\right)^2$$

B 14

$$Q_{i,j} = \frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left\{\log\frac{p_i(k_n|x_n)}{p_i(k|x_n)}\right\}\left\{\log\frac{p_j(k_n|x_n)}{p_j(k|x_n)}\right\}$$

B 15

$$P_i = \frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\mathsf{E}(k_n,k)\left\{\log\frac{p_i(k_n|x_n)}{p_i(k|x_n)}\right\}$$

B 16

$$\frac{1}{\sum_{n=1}^{N}L_n}\sum_{n=1}^{N}\sum_{k\neq k_n}\mathsf{E}(k,k_n)\cdot S\left(\log\frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)}\right)$$

B 17

$$\sum_{j=1}^{M} \lambda_j = 1$$

B 18

$$Q'_{i,j} = \frac{1}{(K-1)N} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathsf{E}(k, k_n) \left\{ \log \frac{p_i(k_n|x_n)}{p_i(k|x_n)} \right\} \left\{ \log \frac{p_j(k_n|x_n)}{p_j(k|x_n)} \right\}$$

B 19

$$P'_0 = 1$$

$$P'_i = \frac{B}{(K-1)N} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathsf{E}(k, k_n) \left\{ \log \frac{p_i(k_n|x_n)}{p_i(k|x_n)} \right\}$$

B 20

CLAIMS:

1. A method for recognizing a pattern that comprises a set of physical stimuli, said method comprising the steps of:

- providing a set of training observations and through applying a plurality of association models ascertaining various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning a particular training observation to one or more associated pattern classes;
- setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and joining thereto a normalization quantity to produce a compound association distribution;
- optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution;
- recognizing target observations representing a target pattern with the help of said compound distribution.

2. A method for modelling an association distribution for patterns that comprise a plurality of physical stimuli; said method comprising the steps of:

- providing a set of training observations and through applying a plurality of association models ascertaining various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning a particular training observation to one or more associated pattern classes;
- setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and joining thereto a normalization quantity to produce a compound association distribution;
- optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution.

3. A method for recognizing a pattern that comprises a set of physical stimuli, said method comprising the steps of:

- receiving a plurality of association models indicating various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning an observation to one or more associated pattern classes, as being combined in a log/linear association distribution according to respective

weight factors, and joined thereto a normalization quantity to produce a compound association distribution;

- optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution;
- recognizing target observations representing a target pattern with the help of said compound distribution.

4. A method as claimed in Claim 1, wherein said association model is a probability model, and said association distribution is a probability model for associating.

5. A method as claimed in Claim 1, wherein said optimizing is effected through minimizing a training error rate in an iterative manner, wherein said error rate is expressed in a continuous and differentiable manner as a function of rivalry values of non-optimum assigning.

6. A method as claimed in Claim 1, wherein said optimizing is effected in a least squares method between an actual discriminant function as resulting from said compound distribution and an ideal discriminant function, as expressed on the basis of an error rate, whilst expressing the weight vector Λ in a closed expression as $\Lambda = Q^{-1}P$, wherein:

Q: autocorrelation matrix of the discriminant functions of the various models

P: correlation vector between the error rate and the discriminant functions.

7. A method as claimed in Claim 6, wherein the empirical error rate is smoothed through representing it as a second degree curve in an interval (-B,A), whilst normalizing Λ through a constraining $\sum \lambda_j = 1$, whilst furthermore expressing the weight vector Λ in a closed expression according to $\Lambda = Q'^{-1}P'$, wherein Q' is an extended autocorrelation matrix through a normalization item added, and P an extended correlation vector through a further normalization item added.

8. A method as claimed in Claim 4, and applied to speech recognition, wherein said probability models comprise one or more of the set of:
as language models: bigram, trigram, fourgram,

as acoustic models: word-internal triphones, cross-word triphones, maximum likelihood linear regression adaptation models;

as additional models: unigram, distance-1-bigram (the middle element being assumed don't care), pentaphones.

9. A system being arranged for executing a method as claimed in Claim 1 for recognizing a pattern that comprises a set of physical stimuli, said system comprising:

- pickup means for receiving a body of training observations and being interconnected to first processing means for through a plurality of stored association models ascertaining various measuring values $p_j(k \mid x)$, $j=1...M$, that each pertain to assigning a particular observation to one or more classes of patterns;

- second processing means fed by said first processing means and being arranged for setting up a log/linear association distribution by combining all association models of the plurality according to respective weight factors, and for joining thereto a normalization quantity to produce a compound association distribution;

- third processing means fed by said second processing means for optimizing said weight factors for thereby minimizing a detected error rate of the actual assigning to said compound distribution;

- recognizing means fed by said third processing means for recognizing target observations representing a target pattern with the help of said compound distribution.

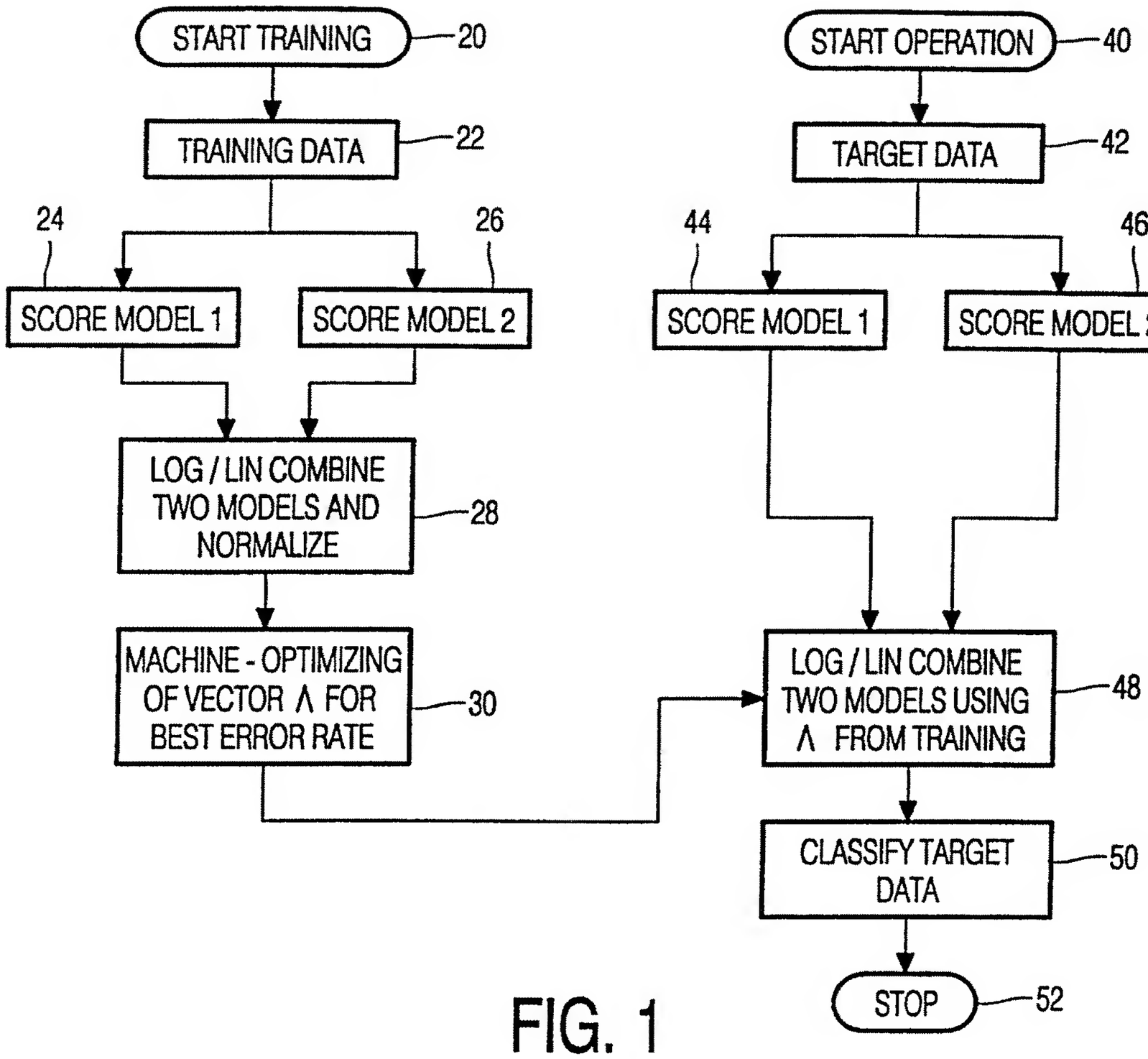

START TRAINING
20
TRAINING DATA
22
24
26
SCORE MODEL 1
SCORE MODEL 2
LOG / LIN COMBINE TWO MODELS AND NORMALIZE
28
MACHINE - OPTIMIZING OF VECTOR Λ FOR BEST ERROR RATE
30
START OPERATION
40
TARGET DATA
42
44
46
SCORE MODEL 1
SCORE MODEL 2
LOG / LIN COMBINE TWO MODELS USING Λ FROM TRAINING
48
CLASSIFY TARGET DATA
50
STOP
52

FIG. 1

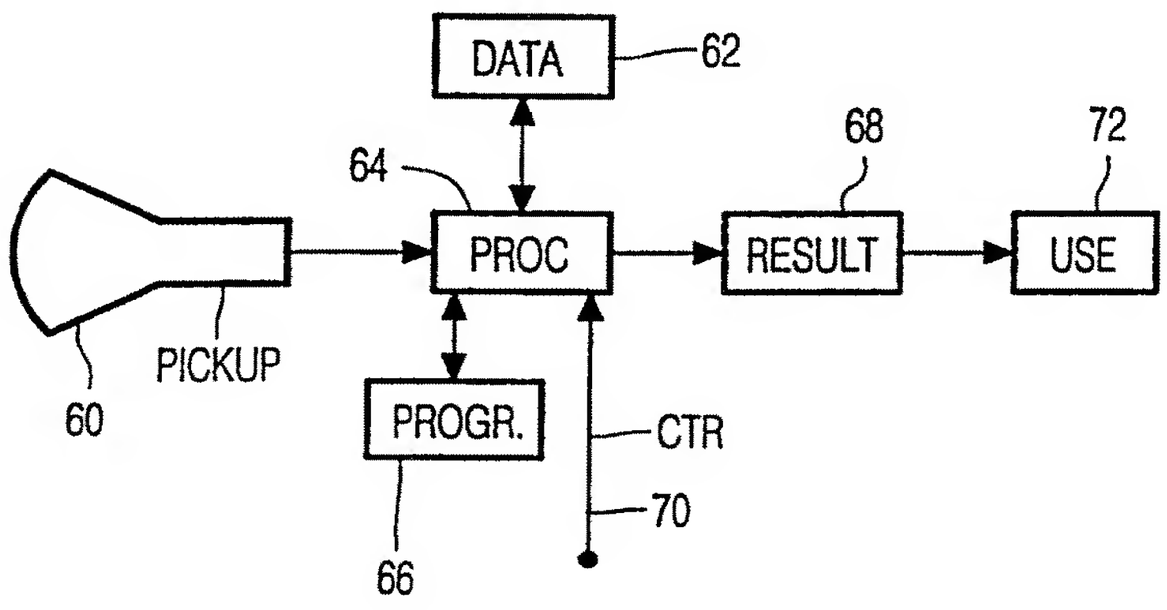

DATA
62
64
PROC
68
RESULT
72
USE
PICKUP
60
PROGR.
66
CTR
70

FIG. 2

(19)日本国特許庁（ＪＰ）　(12)**公表特許公報**（Ａ）　(11)特許出願公表番号
特表2001－511267
（P2001－511267A）
(43)公表日　平成13年８月７日(2001.8.7)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　15/10		Ｇ１０Ｌ　3/00	５３１Ｆ	
15/18			５３７Ｄ	
			５３１Ｇ	

審査請求　未請求　　予備審査請求　有　　（全 27 頁）

(21)出願番号　特願平11－532254
(86)(22)出願日　平成10年12月11日(1998.12.11)
(85)翻訳文提出日　平成11年８月11日(1999.8.11)
(86)国際出願番号　ＰＣＴ／ＩＢ９８／０１９９０
(87)国際公開番号　ＷＯ９９／３１６５４
(87)国際公開日　平成11年６月24日(1999.6.24)
(31)優先権主張番号　１９７５５１９１．２
(32)優先日　平成９年12月12日(1997.12.12)
(33)優先権主張国　ドイツ（ＤＥ）
(31)優先権主張番号　９８２０３７２５．１
(32)優先日　平成10年11月６日(1998.11.6)
(33)優先権主張国　ヨーロッパ特許庁（ＥＰ）

(71)出願人　コーニンクレッカ　フィリップス　エレクトロニクス　エヌ　ヴィ
オランダ国　5621　ベーアー　アインドーフェン　フルーネヴァウツウェッハ　1
(72)発明者　ベヤーレイン，ペーター
オランダ国，5656　アーアー　アインドーフェン　プロフ・ホルストラーン　6
(74)代理人　弁理士　伊東　忠彦　（外１名）

最終頁に続く

(54)【発明の名称】　音声パターン認識用のモデル特殊因子の決定方法

(57)【要約】
物理的刺激の組により構成されるパターンを認識する本発明の方法は、１組の学習用観測量を供給し、複数の連合モデルを適用することにより、特定の学習用観測量の一つ以上の連合したパターンクラスへの割当に関連した種々の測定値ｐｊ（ｋ｜ｘ），ｊ＝１．．．Ｍを確定する段階と、複数の連合モデルを夫々の重み係数に応じて全て結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、その対数／線形連合分布に正規化量を併合する段階と、上記複合分布への実際の割当について検出される誤り率を最小限に抑えるため上記重み係数を最適化する段階と、上記複合分布を用いてターゲットパターンを表現するターゲット観測量を認識する段階とを含む。

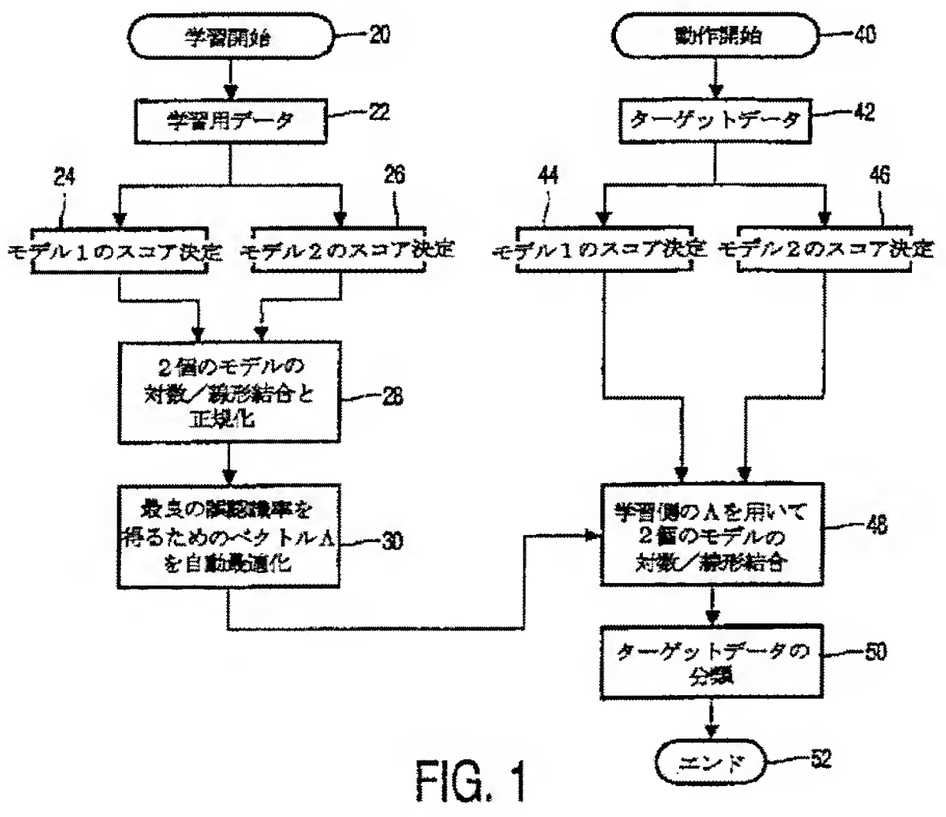


FIG. 1

【特許請求の範囲】

１．　物理的刺激の組により構成されたパターンを認識する方法において、

１組の学習用観測量を供給し、複数の連合モデルを適用することにより、特定の学習用観測量の一つ以上の連合したパターンクラスへの割当に関連した種々の測定値$p_j(k|x)$，$j=1...M$を確定する段階と、

複数の連合モデルを夫々の重み係数に応じて全て結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、その対数／線形連合分布に正規化量を併合する段階と、

上記複合分布への実際の割当について検出される誤り率を最小限に抑えるため上記重み係数を最適化する段階と、

上記複合分布を用いてターゲットパターンを表現するターゲット観測量を認識する段階とを含む方法。

２．　複数の物理的刺激により構成されたパターンの連合分布をモデリングする方法において、

１組の学習用観測量を供給し、複数の連合モデルを適用することにより、特定の学習用観測量の一つ以上の連合したパターンクラスへの割当に関連した種々の測定値$p_j(k|x)$，$j=1...M$を確定する段階と、

複数の連合モデルを夫々の重み係数に応じて全て結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、その対数／線形連合分布に正規化量を併合する段階と、

上記複合分布への実際の割当について検出される誤り率を最小限に抑えるため上記重み係数を最適化する段階とを含む方法。

３．　物理的刺激の組により構成されたパターンを認識する方法に

おいて、

一つ以上の連合したパターンクラスへの割当に関連した種々の測定値$p_j(k|x)$，$j=1...M$を示す複数の連合モデルであって、夫々の重み係数に応じて対数／線形連合分布に結合され、複合連合分布を生成するため正規化量が併合される複数の連合モデルを受信する段階と、

上記複合分布への実際の割当について検出される誤り率を最小限に抑えるため上記重み係数を最適化する段階と、

上記複合分布を用いてターゲットパターンを表現するターゲット観測量を認識する段階とを含む方法。

４．　上記連合モデルは確率モデルであり、上記連合分布は連合用の確率モデルである、請求項１記載の方法。

５．　最適化は反復的な形で学習誤り率を最小化することにより実現され、上記誤り率は非最適割当の対抗値の関数として連続微分可能な形式で表現される、請求項１記載の方法。

６．　最適化は、誤り率に基づいて表現されるような上記複合分布の結果として得られる実際の識別関数と理想的な識別関数との間で最小二乗法を用いて実現され、

Ｑが種々のモデルの識別関数の自己相関マトリックスを表し、Ｐが上記誤り率と上記識別関数との間の相関ベクトルを表すときに、

重みベクトルΛは、

$$\Lambda = Q^{-1} P$$

のように閉じた形式で表現される、請求項１記載の方法。

７．　経験的な誤り率は、間隔（－Ｂ，Ａ）内で２次曲線として表現することにより平滑化され、

上記重みベクトルΛは

$$\Sigma \lambda_i = 1$$

という形で拘束され、

Ｑ’が正規化項を付加することにより拡張された自己相関マトリックスを表し、Ｐ’が別の正規化項を付加することにより拡張された自己相関ベクトルを表すときに、

上記重みベクトルは、

$$\Lambda = Q'^{-1} P'$$

に従って閉じた形式で表現される、請求項６記載の方法。

８．　音声認識に適用され、

上記確率モデルは、

言語モデルとしての２－ｇｒａｍ、３－ｇｒａｍ、４－ｇｒａｍの組と、

音響モデルとしてのワード・インターナル・トライフォン、クロス・ワード・トライフォン、最尤線形回帰アダプテーションモデルの組と、

付加的なモデルとしての１－ｇｒａｍ、中間要素はドントケアであると考えられる距離１の２－ｇｒａｍ、ペンタフォンの組の中の一つ以上の組を含む、請求項４記載の方法。

９．　記憶された複数の連合モデルを用いて、特定の学習用観測量の一つ以上のパターンクラスへの割当に関連した種々の測定値ｐ」（ｋ｜ｘ），ｊ＝１．．．Ｍを確定する第１の処理手段に相互接続され、学習用観測量の本体を受信するピックアップ手段と、

上記第１の処理手段の下流に接続され、それぞれの重み係数に従って上記複数の連合モデルをすべて結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、正規化量を併合するよう構成された第２の処理手段と、

上記第２の処理手段の下流に接続され、上記複合分布への実際の割当に関して検出された誤り率を最小限に抑えるため上記重み係数を最適化する第３の処理手段と、

上記第３の処理手段の下流に接続され、上記複合分布を用いてターゲットパターンを表現するターゲット観測量を認識する認識手段とを含み、

物理的刺激の組により構成されたパターンを認識する請求項１に記載された方法を実施するシステム。

【発明の詳細な説明】

音声パターン認識用のモデル特殊因子の決定方法

発明の背景

本発明は、物理的刺激の組により構成されたパターンを認識する方法に係わり、この方法は、

１組の学習用観測量を供給し、複数の連合モデルを適用することにより、特定の学習用観測量の一つ以上の連合したパターンクラスへの割当に関連した種々の測定値p_j（k｜x），j＝1．．．Mを確定する段階と、

複数の連合モデルを夫々の重み係数に応じて全て結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、その対数／線形連合分布に正規化量を併合する段階とを含む。

本発明は音声認識を想定しているが、音声理解、音声翻訳、並びに、手書き文字認識、顔の認識、情景認識、及び、その他の環境の認識のような他の認識プロセスにも同じように適用可能である。連合モデルは、パターンをクラスに割り当てる確率分布を与える確率モデルである。他のモデルは、ファジー論理、或いは、ターゲットとクラスとの間で測定された距離のような類似した測度に基づく。従来の技術は、合成された認識の取り組みにおいてかかる種々のモデルを使用するが、種々の協働するモデルに与えられる影響は偶然的に決まる。これは、僅かな基本モデル及び／又は唯一の基本モデルだけが実施できることを意味する。

本願の発明者は、最大エントロピー原理及び識別型学習原理の統合は、二つ以上のモデルを合成する場合に、原則として、従来のヒューリスティックな方法よりも優れた結果が得られることを見出した。また、直接的なデータ処理手続は、これらの結果を低コストかつ高速に与える。

したがって、特に、本発明の目的は、識別型学習を通じてパラメータ推定を行いながら、種々のサブモデルp_j（k｜X）の対数－線形結合を推定することである。以下、本発明の一面によれば、、物理的刺激の組を含むパターンを認識することにより上記本発明の目的を達成する方法は、

１組の学習用観測量を供給し、複数の連合モデルを適用することにより、特定

の学習用観測量の一つ以上の連合したパターンクラスへの割当に関連した種々の測定値ｐ」（ｋ｜ｘ），　ｊ＝１．．．Ｍを確定する段階と、

複数の連合モデルを夫々の重み係数に応じて全て結合することにより対数／線形連合分布を設定し、複合連合分布を生成するため、その対数／線形連合分布に正規化量を併合する段階と、

上記複合分布への実際の割当の検出される誤り率を最小限に抑えるため上記重み係数を最適化する段階と、

上記複合分布を用いてターゲットパターンを表現するターゲット観測量を認識する段階とを含む。特に、このような処理によって、任意の数のモデルを単一の最大エントロピー分布に合成できるようになる。また、特性及び表現が非常に広範囲に変化するモデルの相互作用を最適化することができる。

また、本発明は、上記本発明による連合分布をモデル化する方法に関する。これにより、複合分布の以降のユーザが適切なパターンを認識するための優れたツールが得られる。

また、本発明は、本発明によって生成された連合分布を用いてパターンを認識する方法に関する。この方法は、上記本発明によって実現されたツールを適用することによりユーザに多大の利益を供与する。

本発明は、上記本発明による方法を実施するため構成されたシステムに関する。本発明の更なる局面は従属した請求項に記載されている。

図面の簡単な説明

以下、好ましい実施例の詳細な説明を、特に、添付図面と共に参照して、本発明の上記並びに他の局面及び利点について詳細に説明する。図面中、

図１は、本発明の方法の全体的なフローチャートであり、

図２は、本発明を実施する統合システムの構成図であり、

図３乃至２１には、本発明による自動化方法と共に使用される種々の数式Ｂ１－Ｂ２０が表されている。

好ましい実施例の詳細な説明

本発明は、非常に多様な特性を有する物理量の取扱及び調節に関する数学のバ

ランスのとれたアプリケーションに基づき、本発明の開示の大部分は高等数学に基づく。しかし、スタート点と最終的な結果の両方は、不変的な物理的な面を有し、関連性がある。音声認識は、種々のタイプの機械を制御するため使用される。情景解析は無人自動車を誘導する。画像認識はゲートの開閉制御に使用される。これら以外にも種々のアプリケーションがある。以下では、数式は通し番号が付けられ、本文中で数式はその番号によって参照される。

本発明は、特に、音声認識用の幾つかの異なるモデルを単一のパターン認識処理に合成し、最適化するため、モデル特殊因子を決定する。

統計的な音声認識モデルは、最小誤識別率の識別メカニズムを形成するためベイズ判定理論を利用する。この理論に従って、所定の事後又は「現実の」確率分布π（ｋ｜ｘ）に対し、

$$\forall k' = 1,\ldots,K;\ k' \neq k : \log \frac{\pi(k|x)}{\pi(k'|x)} \geq 0 \Rightarrow x \in k \qquad (1)$$

が成立するときに、観測量ｘがクラスｋに割り当てられる（簡単のため、ｘ∈ｋ）ような判定が行われる。

用語上、項ｌｏｇ（π（ｋ｜ｘ）／π（ｋ｜ｘ）は、識別関数と称される。以下、簡単のため、この項は、ｇ（ｘ，ｋ，ｋ’）のように表記する。判定ルール（１）が完全な文（センテンス）を認識するため使用されるとき、時間的な長さＴを有する観測された表現x_1^T＝（x^1，．．．，x^T）は、長さＳの発話語シーケンスw_1^S＝（w^1，．．．，w^S）に分類される。しかし、事後分布π（w_1^S｜x_1^T）は、人の複雑な自然音声会話を記述するので未知である。したがって、これは、分布ｐ（w_1^S｜x_1^T）によって近似されるべきである。ここまでは、パラメトリック確率分布の形式による音声の音響音声学的並びに文法的モデリングは、最良の結果を与える。分布ｐ（w_1^S｜x_1^T）の形式は予め決められ、分布の未知パラメータは学習データに基づいて推定される。かくして獲得された分布ｐ（w_1^S｜x_1^T）は、次に、ベイズ判定に代入される。表現x_1^Tは、

$$\forall w'^{s'}_1 \neq w_1^s : \log \frac{p(w_1^s|x_1^T)}{p(w'^{s'}_1|x_1^T)} > 0 \qquad (2)$$

である語シーケンスw_1^Sに割り当てられる。

識別関数の変換

$$g(s_1^T, w_1^s, w'^{s'}_1) = \log\frac{p(w_1^s|x_1^T)}{p(w'^{s'}_1|x_1^T)} = \log\frac{p(w_1^s)p(x_1^T|w_1^s)}{p(w'^{s'}_1)p(x_1^T|w'^{s'}_1)}, \quad (3)$$

によって、文法モデルｐ（w_1^S）を音響音声学モデルｐ（$x_1^T|w_1^S$）から自然な形で分離することができる。文法モデルｐ

（w_1^S）は、特に語シーケンスw_1^Sの出現確率を記述し、音響音声学モデルｐ（$x_1^T|w_1^S$）は、語シーケンスw_1^Sの発話中の音響信号x_1^Tの出現確率を推定する。両方のモデルは別々に推定できるので、比較的限定された個数の学習データが最適に使用される。判定ルール（３）は、たとえ、分布ｐの推定が最適であったとしても、分布ｐの未知分布πからの偏差に起因して最適に達しない場合がある。これにより、所謂識別法の使用が促される。識別型法は、学習データに基づいて経験的に測定された判定ルールの誤識別率に関して分布ｐを直接的に最適化する。このような識別型最適化の最も簡単な例は、所謂言語モデル因子λを使用することである。このとき、式（３）は以下の通り変形される。

$$g(x_1^T, w_1^s, w'^{s'}_1) = \log\frac{p(w_1^s)^\lambda p(x_1^T|w_1^s)}{p(w'^{s'}_1)^\lambda p(x_1^T|w'^{s'}_1)} \quad (4)$$

実験的に、判定ルール（４）によって生ずる誤識別率は、λ＞１となるようλを選択するときに減少する。λ＝１である理論値から外れる理由は、複合事象（w_1^S，x_1^T）の確率のモデル化が不完全若しくは不正確である点にある。事象（w_1^S，x_1^T）を発生させる過程についての知識は不完全であるため、モデリングの不正確さは回避できない。

従来、多数の音響音声学的並びに文法的言語モデルが解析されている。これらの解析の目的は、既知若しくは所与のモデルの中から着目している認識作業に対

する「最良」モデルを見つけることである。このような方式で決定されたすべてのモデルは、実際の確率分布の不完全な表現であるので、これらのモデルが音声認識のようなパターン認識に使用されるとき、クラスへの誤った割当として誤認識が生ずる。

本発明の目的は、現実の確率分布をより厳密に近似し、僅かな処理労力を加えるだけで実施することができ、特に、より多数の既知若しくは所与のモデルを単一の分類メカニズムに簡単に統合できるモデリング、特に音声用モデリングを提供することである。

発明の概要

本発明の解決法の新規な局面は、既知の音声特性を、複雑かつ困難な学習を伴う単一の音響音声学的分布モデルや単一の文法的分布モデルに統合しようとしない点である。多種多様な音声音響学的並びに文法的特性が、種々の分布$p_j(w_1^S \mid x_1^T)$，$j=1...M$の形式で別々にモデル化され学習され、次に、複合分布

$$p^{\pi}_{\{\Lambda\}} = (w_1^S \mid x_1^T) = C(\Lambda).\prod_{j=1}^{M} p_j(w_1^S \mid x_1^T)^{\lambda_j} = \exp\left\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log p_j(w_1^S \mid x_1^T)\right\} \quad (5)$$

に統合される。モデルp_jの分布$p^{\pi}_{\{\Lambda\}}$に対する影響は、関連した係数λ_jよって決定される。

因子C（Λ）は、確率に対する正規化条件が満たされたこと保証する。自由因子$\Lambda=(\lambda_1, ..., \lambda_M)^{tr}$は合成された識別関数

$$g(x_1^T, w_1^S, w'^{S'}_1) = \log \frac{\prod_{j=1}^{M} p_j(w_1^S \mid x_1^T)^{\lambda_j}}{\prod_{j=1}^{M} p_j(w'^{S'}_1 \mid x_1^T)^{\lambda_j}} \quad (6)$$

の誤識別率ができる限り小さくなるように調整される。この基本的なアイデアを実現するために多数の方法が考えられるが、以下では、その中の幾つかの方法を説明する。

最初に、以下の説明で使用される種々の用語を定義する。各語シーケンスw_l^Sはクラスｋを形成し、シーケンス長Ｓはクラス毎に異なる。音声発話x_l^Tは観測量ｘであると考えられ、その長さＴは、観測量毎に異なる。

画像データは、（x_n，ｋ）によって表記され、ｎ＝１，．．．，Ｎ；ｋ＝０，．．．，Ｋである。尚、Ｎは音響学習観測量x_nの個数であり、k_nは観測量x_nと関連した正しいクラスを表す。また、$k \neq k_n$なるｋは、k_nに関して競合する種々の不正確な対抗クラスである。

ベイズ判定ルール（１）に従って観測量ｘをクラスｋに分類する場合を考える。観測量ｘはクラスｋの音響的な実現形式である。音声認識の場合に、各クラスｋは語のシーケンスを代表する。しかし、この方法はより汎用的に適用することが可能である。

学習用観測量x_nによって生成されるクラスk_nは既知であり

、理想的な経験に基づく分布$\hat{\pi}$（ｋ｜ｋ）は、学習データ（x_n，

ｋ）；ｎ＝１．．．Ｎ；ｋ＝０．．．Ｋに基づいて構築され得る。この分布は、この分布から獲得された判定ルールが学習データに適用された場合に最小の誤識別率を有するように作られるべきである。完全な語シーケンスｋを分類する場合に、誤りのある語シーケンス$k \neq k_n$を選択することによる分類誤りは、幾つかの語の誤りを生じさせる。正しくないクラスｋと正しいクラスk_nとの間の語の誤りの数は、レーベンシュタインの距離Ｅ（ｋ，k_n）と呼ばれる。Ｅ（ｋ，k_n）から形成される判定ルールは、単調特性が満たされるとき、最小の語誤識別率を有する。

理想的な経験に基づく分布$\hat{\pi}$は、学習データに対してだけ与えら

れ、正しいクラス割当てが得られない未知のテストデータに対しては定義されな

い経験的誤り値Ｅ（ｋ，ｋｎ）の関数である。したがって、この分布に基づいて、任意の独立したテストデータに対し定義され、学習データに関する経験的な誤識別率ができる限り小さ

く抑えられた分布

$$p^{\pi}\{\Lambda\}(k|x) = \frac{\exp\left\{\sum_{j=1}^{M} \lambda_j \log p_j(k|x)\right\}}{\sum_{k'=1}^{K} \exp\left\{\sum_{j=1}^{M} \lambda_j \log p_j(k'|x)\right\}} \qquad (7)$$

が求められる。M個の予め決められた分布モデルｐ１（ｋ｜ｘ），．．．．，ｐＭ（ｋ｜ｘ）が任意のテストデータに関して定義

されるとき、上記式は分布$p^{\pi}_{\{\Lambda\}}$（ｋ｜ｘ）に対しても成立する。自由

に選択可能な係数Λ＝（λ１，．．．，λＭ）^{tr}が、学習データに

関する$p^{\pi}_{\{\Lambda\}}$（ｋ｜ｘ）の誤りを最小に抑えられるように決定され、

学習データが典型的である場合に、$p^{\pi}_{\{\Lambda\}}$（ｋ｜ｘ）は独立したテス

トデータに関して最適な判定を行う必要がある。

ＧＰＤ法並びに最小二乗法は、分類器の平均誤識別率を近似する規準を最適化する。ＧＰＤ法と比べて、最小二乗法は最適な係数Λに対し閉じた解を与える点が有利である。

次に、最小二乗法の場合について考える。識別関数（１）は分類器の品質を決定するので、係数Λは経験的な誤識別率Ｅ（ｋ，ｋ。

）からの分布$p^{\pi}_{\{\Lambda\}}$（ｋ｜ｘ）の識別関数の平均二乗根偏差Ｂ１４を

最小限に抑える必要がある。ｒに関する加算は規準のすべての対抗クラスを含む。Ｄ（Λ）を最小化することにより、式Ｂ１５及びＢ１６により詳細に示されている最適な係数ベクトル

$$\Lambda = Q^{-1} P \quad (9)$$

に対する閉形式の閉じた解が得られる。

式中、Ｑは所定の分布モデルの識別関数の自己相関マトリックス

である。ベクトルＰは、所定のモデルの識別関数と分布$\hat{\pi}$の識別関

数との間の関係を表現する。

仮説ｋの語の誤り率Ｅ（ｋ，ｋ$_n$）は係数λ_1，．．．，λ_M内で線形に選択される。逆に、分布モデルｐ$_i$の識別能力は、識別関数

$$\log \frac{p_i(k|x_n)}{p_i(k_n|x_n)}$$

によって直接的に係数を決めるため、係数λ_1，．．．，λ_M内に線形に包含される。

或いは、これらの係数はＧＰＤ法を用いて決定してもよい。ＧＰＤ法の場合に、平滑化された経験に基づく以下の誤り率Ｅ（Λ）は学習データに対し直接的に平滑化され得る。

$$E(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} l(x_n, k_n, \Lambda) \quad (12)$$

$$l(x_n, k_{nO}, \Lambda) = \left\{ 1 + A \left\{ \frac{1}{K} \sum_{r=1}^{K} \exp\left\{ -\eta \log \frac{p^{\pi}_{\{\Lambda\}}(k_n|x_n)}{p^{\pi}_{\{\Lambda\}}(k|x_n)} \right\} \right\}^{-\frac{B}{\eta}} \right\}^{-1} \quad (13)$$

左辺の式は、観測量ｘ$_n$を誤って分類する危険に対する平滑化された測定量である。値Ａ＞０，Ｂ＞０，η＞０は、誤り分類危険の平滑化のタイプを決定し、予め適切に与えられるべきである。Ｅ（λ）が対数線形結合の係数λに関して最小化されるとき、λ$_i$に対し、ステップ幅Ｍを有する反復式

$$\lambda_j^{(0)} = 1 \qquad (11)$$

但し、ｊ＝１，．．．．，Ｍ

が得られる。また、式Ｂ１３及びＢ１４に従って、

$\Lambda^{(1)} = (\lambda_1^{(1)}, \ldots, \lambda_M^{(1)})^{tr}$；ｊ＝１，．．．．，Ｍが得られる。

係数ベクトルΛは、識別関数

$$\log \frac{p^{\pi}_{\{\Lambda\}}(k_n|x_n)}{p^{\pi}_{\{\Lambda\}}(k|x_n)} \qquad (12)$$

を用いて、規準Ｅ（Λ）に包含されることに注意する必要がある。

仮にＥ（Λ）が減少するとき、識別関数（１２）は、式（９）及び（１０）のため、平均的に増加する。この結果として、判定ルールは更に改良される（式（１）を参照のこと）。

上記説明では、すべての利用可能な知識源を単一のパターン認識システムに統合することが目的であり、二つの原理が併合される。第１の原理は最大エントロピー原理である。この原理は、導入される仮説をできる限り減らし、その結果として不確実さを最大化するように作用する。そのため、指数関数的な分布が使用される。この方法では、知識源の組合せの構造が定義される。第２の原理は、種々の知識源に割り当てられる重み付け係数及び関連したモデルを決めるため、識別型学習を行う。パラメータを最適化することにより、誤りは最小限に抑えられる。音声の場合、モデルは、意味論的モデル、統語論的モデル、音響的モデル及びその他のモデルなどである。

この方法は、多種のサブモデルを対数線形結合し、識別型学習を通じてパラメータを推定する。このようにして、サブモデルの追加は認識スコアを改善させる。さもなければ、着目中のモデルは無視される。しかし、サブモデルは決して認識精度を低下させない。か

くして、すべての利用可能なサブモデルが最適な結果を生ずるため合成される。本発明の別のアプリケーションは、既存のモデル合成を新しい認識環境に適応させることである。

この処理の理論的なアプローチには以下の様々なステップが含まれる。

－経験的誤認識率のパラボリック平滑化

－「最小誤認識率学習」の理論の簡単化

－反復シーケンスを必要としない閉じた形式の解を供与すること　また、本発明によれば、以下の付加的な機能が付与される。

－最適言語モデル因子の推定

－対数線形隠れマルコフモデルの適用

－適モデル合成のための閉じた形式の式

－クラス特殊確率分布の識別型学習のための閉じた形式の式

以下、式（１）に指定された分類作業のため、真又は事後分布π（ｋ｜ｘ）は未知であるが、モデル分布ｐ（ｋ｜ｘ）によって近似される。二つの分布は、不正確なモデリング仮説と、不十分なデータとに起因して異なる。その一例は、式Ｂ１に使用される言語モデル因子λである。

形式的な定義は、式（５）に与えられるように種々のサブモデル結合に続いて、項ｌｏｇＣ（Λ）は、形式的な確率分布を得るため正規化を行う。これにより得られた識別関数は、

$$\log\frac{P_{\{\Lambda\}}(k|x)}{P_{\{\Lambda\}}(k'|x)} = \sum_j \lambda_j \log\frac{p_j(k|x)}{p_j(k'|x)} \qquad (B\,2)$$

である。

誤認識率は最小化され、Λは最適化される。文レベルの最適化は

以下の通り行われる。

・クラスｋ：語シーケンス

・観測量ｘ：発話（例えば、文）

・正しい文を与えるＮ個の学習サンプルx_n

・各サンプルx_nに対し、

－k_n：会話として正しいクラス

－$k \neq k_n$：起こり得るすべての文、或いは、例えば、その妥当な部分集合で

ある対抗クラス

・クラスの類似性：Ｅ（k_n，ｋ）

－Ｅ：レーベンシュタイン距離、或いは、単調である同等に適切な測定量の適当な関数

・語シーケンスk_n内の語数：L_n

次に、式Ｂ３は、目的関数である経験的誤認識率を与える。式中、左辺は、クラスｋとk_nの間の誤りのある偏差の数に基づく最尤クラスを導く。

パラメータΛは以下のように推定される。

・反復的な解を与える一般化された確率的降下法（ＧＰＤ）による最小誤識別率学習

・パラボリック（双曲線型）平滑化と組み合わされ、閉じた形式の解を与える最小誤識別学習の変形

・閉じた形式の解を与える最小二乗法に基づく第３の方法

ＧＰＤ法の場合に、平滑化された経験的誤識別率最小化は式Ｂ４に基づく。平滑化された誤分類危険は式Ｂ５によって与えられ、平均的な競争は式Ｂ６により与えられる。

平滑化された経験的誤識別率は式Ｂ７によって最小化される。式中、１は、簡便な計算の場合に微分可能でなければならない損失関数である。競争は式Ｂ８によって与えられ、式中、Ｅは誤りの数を示す。平均競争は、式Ｂ９において加算することにより与えられる。平滑化された誤分類危険は、シグモイド関数のような挙動を示す式

Ｂ１０によって表現される。$R_n=-\infty$の場合に、損失関数１は零になり、$R_n=+\infty$の場合に、限界値は１＝１である。式中、Ａ，Ｂは零よりも大きいスケーリング定数である。Λに関する微分によって、式Ｂ１１が得られ、ここで、ベクトル$\Lambda^{(1)}$は式Ｂ１２によって与えられ、最後の結果は式Ｂ１３によって与えられる。

また、本発明は、識別型モデル合成ＤＭＣを見つけるための閉じた形式の解を提供する。この解は、最小二乗法に従って、識別関数と理想的な識別関数Ｅ（ｋ

$_n$，ｋ）との間の距離を最小化させる。基本的な式はＢ１４に示される。ここで、$\Lambda = Q^{-1} P$であり、式中、Ｑは式Ｂ１５で表された要素$Q_{i,j}$を有するマトリックスである。また、Ｐは式Ｂ１６で表された要素P_iを有するベクトルである。経験的誤識別率は既にＢ３に記載されている。計算上の理由から、経験的誤識別率は、式Ｂ１７によって表現されるような平滑化された経験的誤識別率によって近似される。ここで、ｋとk_nとの間の誤りの数が、シグモイド関数Ｓ又は同様に有効な関数を用いて表される。有効な形式は、$S(x) = \{(x+B)/(A+B)\}^2$であり、式中、$-B < x < A$かつ$-B < 0 < A$である。ｘの値に対し、Ｓ＝１であり、小さいｘの値に対し、Ｓ＝０である。このパラボラは有効であることが分かった。種々の他の二次曲線が有効であることが判明した。関連した対抗側は、Ｓの中心及びパラボラ的に湾曲した間隔に存在する必要がある。次に、最終的に、正規化定数が式Ｂ１８に従ってΛに対し加算される。

第２の規準は、マトリックス計算式$(\alpha, \lambda^{tr})^{tr} = Q'^{-1} P'$に従って解法され、ここで、$Q'_{0,0} = 0$、$Q'_{0,j} = 1$及び$Q'_{j,0} = 1/2(A+B)^2$に従って付加的な行及び列が正規化のためマトリックスＱ’に付加される。相関マトリックスＱ’の一般的な要素は式Ｂ１９に与えられる。閉じた解は平滑化ステップ関数ｓによって実現可能にされることに注意する必要がある。また、ベクトルＰ’は、同様に正規化用要素$P'_0 = 1$が与えられ、一方、ベ

クトルＰ’の一般的な要素は式Ｂ２０に与えられる。

２－ｇｒａｍ、３－ｇｒａｍ、４－ｇｒａｍ又は５－ｇｒａｍモデルのような多種のＭ－ｇｒａｍ言語モデルや、ワード・インターナル・トライフォン(word-internal triphones)、クロス・ワード・３－ｇｒａｍ(cross-word trigram)及びペンタフォン(pentaphones)モデルのような種々の音響モデルを用いて実験が行われる。一般的に、自動ＤＭＣ処理は、同じサブモデルの集合を用いて非自動精密チューニングによって生成される結果と同等に優れた結果を実現する。しかし、本発明の自動処理による付加的なサブモデルの追加は、誤りの数を約８％減少させることができる。これは、改良された音声認識の技術における重大な前進であると考えられる。本発明は、適切なサブモデルが利用できるならば、署名、

手書き文字、情景解析などの別のタイプのパターンを認識するため同様に優れた結果を与えるものと期待される。一般的な認識のため使用される他のサブモデルには、ｍｌｌｒアダプテーション、１－グラム(unigram)、中間要素はドントケアであるとみなされる距離１の２－グラム(distance-1-bigram)、ペンタフォン(pentaphones)及びｗｓｊモデルが含まれる。このような環境で、本発明の自動処理におけるサブモデルの数を増加させることにより、誤りの個数は８～１３％の有意な量が減少される。

図１には、本発明の方法の全体的なフローチャートが示されている。ブロック２０では学習が開始され、学習用データ又はパターンはブロック２２で与えられる。始めに、要求されるソフトウェア及びハードウェア、特に、サブモデルが必要に応じて宣言され、多様なパターンの識別が行われる。簡単のため、サブモデルの個数は２個に制限されている場合を考えるが、サブモデルの個数は３個以上でも構わない。並行したブロック２４及び２６において、個々のサブモデルに対するスコアが決定される。ブロック２８において、種々のサブモデルの対数線形結合が行われ、正規化される。ブロッ

ク３０において、最小の誤認識率が達成されるという観点でベクトルΛの自動最適化が行われる。尚、ベクトルΛは、関連したサブモデル若しくはモデルが全く改良を行わないことを知らせるため１個以上の零値成分を有することに注意する必要がある。

次に、図１の右側に示されるようにベクトルΛ及び種々の適用可能なサブモデルがターゲットデータを認識するため使用される。左側の学習と右側の運用は、時間的かつ空間的に互いに別々に行われ、例えば、ある人は、プロバイダ側で自分の声に対し機械を学習させる。これには、付加的なデータ処理設備が必要とされる。次に、このように学習された機械は、家庭若しくは車内の環境、又は、それ以外の場所で使用される。したがって、ブロック４０～４６は、同図の左側のブロックと対応する。

ブロック４８において、種々のサブモデルからのスコアが、学習側で見つけられたベクトルΛの種々の成分を用いて対数線形結合される。最後に、ブロック５

０において、ターゲットデータがブロック５０から得られた結果を用いて分類される。ブロック５２において、処理は終了し、準備が完了する。

図２は、本発明を実施するシステムの概略図である。必要な機能は標準的なハードウェア、又は、専用装置上に割り付けられる。ボイスレコーダ、２次元光学式スキャナのような適当なピックアップ６０が、必要に応じてＡ／Ｄ変換機能並びに品質改良前処理と共に設けられる。ブロック６４には、プログラムメモリ６６からのプログラムを、ピックアップ６０から到着したデータ、又は、データ記憶装置６２からのデータに適用する処理が示されている。データ記憶装置６２には、ピックアップ６０から転送されたデータが持続的若しくは一時的に格納される。ライン７０は、スタート／ストップのようなユーザ制御信号、場合によっては、例えば、役に立たないサブモデルを完全に禁止するような学習用補助信号を受信する。

ブロック６８では、例えば、作表、印刷、適切な音声応答を得る

ための会話構造をアドレス指定、或いは、適切な出力制御信号を選択することにより認識結果が使用可能にされる。ブロック７２では、音声応答を出力し、認識された人のためゲートを開き、分類機械内でパスを選択する等の認識された音声の用法が示されている。

$$\log \frac{\pi(x|k) \cdot \pi(k)}{\pi(x|k') \cdot \pi(k')} \quad \rightarrow \quad \log \frac{p(x|k) \cdot p(k)^{\lambda}}{p(x|k') \cdot p(k')^{\lambda}}$$

B1

$$\frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \mathrm{E}\left(\arg\max_{k} \left(\log \frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)} \right), k_n \right) =$$

$$\frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathrm{E}(k_n, k) \cdot \delta\left(k \ , \ \arg\max_{k'} \left(\log \frac{p_{\wedge}(k'|x_n)}{p_{\wedge}(k_n|x_n)} \right) \right)$$

B3

$$L(\wedge) \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, \wedge)$$

B4

$$\ell(x_n, \Lambda) = \frac{1}{1 + AR_n(\Lambda)^{-B}}$$

B5

$$R_n(\Lambda) = \left(\frac{1}{K-1} \sum_{k \neq k_n} \left(e^{E(k_n,k) \cdot \sum_{j=1}^{M} \lambda_j \log \frac{p_j(k|z_n)}{p_j(k_n|z_n)}} \right)^{\eta} \right)^{\frac{1}{\eta}}$$

B6

$$L(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, \Lambda)$$

B7

$$y_{nk}(\Lambda) = \left(\frac{p_{\Lambda}(k|x_n)}{p_{\Lambda}(k_n|x_n)} \right)^{E(k_n,k)}$$

B8

$$R_n(\Lambda) = \left(\frac{1}{K-1} \sum_{k \neq k_n} [y_{nk}(\Lambda)]^{\eta} \right)^{\frac{1}{\eta}} \quad \text{B 9}$$

$$\ell(x_n, \Lambda) = \frac{1}{1 + AR_n(\Lambda)^{-B}} \quad \text{B 10}$$

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} - \varepsilon \sum_{n=1}^{N} \ell(x_n, \Lambda^{(I)}) \left(1 - \ell(x_n, \Lambda^{(I)})\right) \times \frac{\sum_{k \neq k_n} \mathrm{E}(k_n, k) \log \left(\frac{p_j(k|x_n)}{p_j(k_n|x_n)} \right) [y_{nk}(\Lambda^{(I)})]^{\eta}}{\sum_{k \neq k_n} [y_{nk}(\Lambda^{(I)})]^{\eta}} \quad \text{B 11}$$

$$\Lambda^{(I)} = (\lambda_1^{(I)}, \ldots, \lambda_M^{(I)})^{tr} \quad \text{B 12}$$

$$y_{nk}(\Lambda) = \left(\frac{p_{\Lambda}(k|x_n)}{p_{\Lambda}(k_n|x_n)} \right)^{\mathrm{E}(k_n, k)} \quad \text{B 13}$$

$$\frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left(\log\frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)}-\mathsf{E}(k_n,k)\right)^2 =$$

$$\frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left(\sum_j \lambda_j \log\frac{p_j(k|x_n)}{p_j(k_n|x_n)}-\mathsf{E}(k_n,k)\right)^2$$

B 14

$$Q_{i,j} = \frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\left\{\log\frac{p_i(k_n|x_n)}{p_i(k|x_n)}\right\}\left\{\log\frac{p_j(k_n|x_n)}{p_j(k|x_n)}\right\}$$

B 15

$$P_i = \frac{1}{(K-1)N}\sum_{n=1}^{N}\sum_{k\neq k_n}\mathsf{E}(k_n,k)\left\{\log\frac{p_i(k_n|x_n)}{p_i(k|x_n)}\right\}$$

B 16

$$\frac{1}{\sum_{n=1}^{N}L_n}\sum_{n=1}^{N}\sum_{k\neq k_n}\mathsf{E}(k,k_n)\cdot S\left(\log\frac{p_{\wedge}(k|x_n)}{p_{\wedge}(k_n|x_n)}\right)$$

B 17

$$\sum_{j=1}^{M} \lambda_j = 1$$

B 18

$$Q'_{i,j} = \frac{1}{(K-1)N} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathrm{E}(k, k_n) \left\{ \log \frac{p_i(k_n|x_n)}{p_i(k|x_n)} \right\} \left\{ \log \frac{p_j(k_n|x_n)}{p_j(k|x_n)} \right\}$$

B 19

$$P'_0 = 1$$

$$P'_i = \frac{B}{(K-1)N} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathrm{E}(k, k_n) \left\{ \log \frac{p_i(k_n|x_n)}{p_i(k|x_n)} \right\}$$

B 20

【図１】

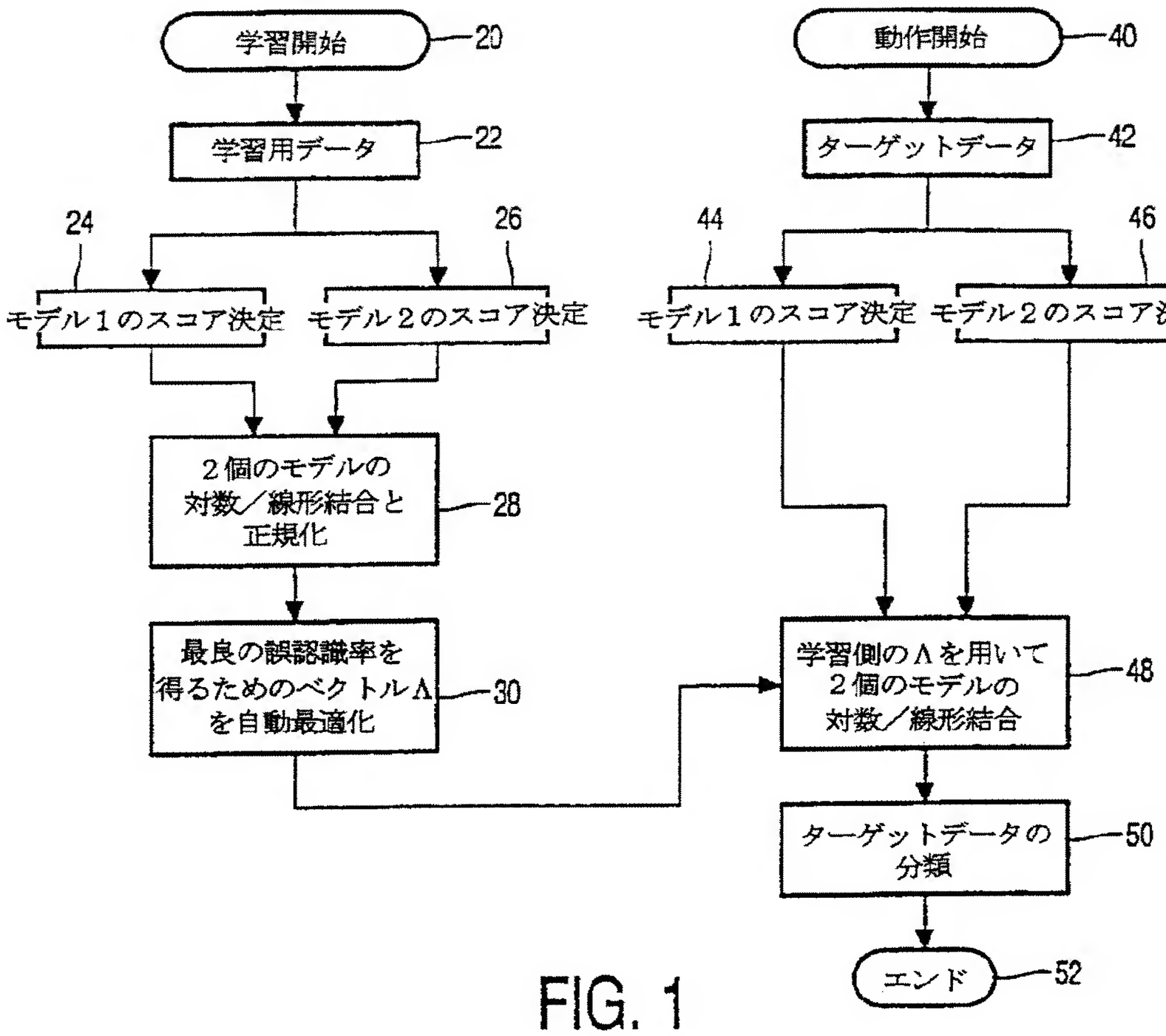


FIG. 1

【図２】

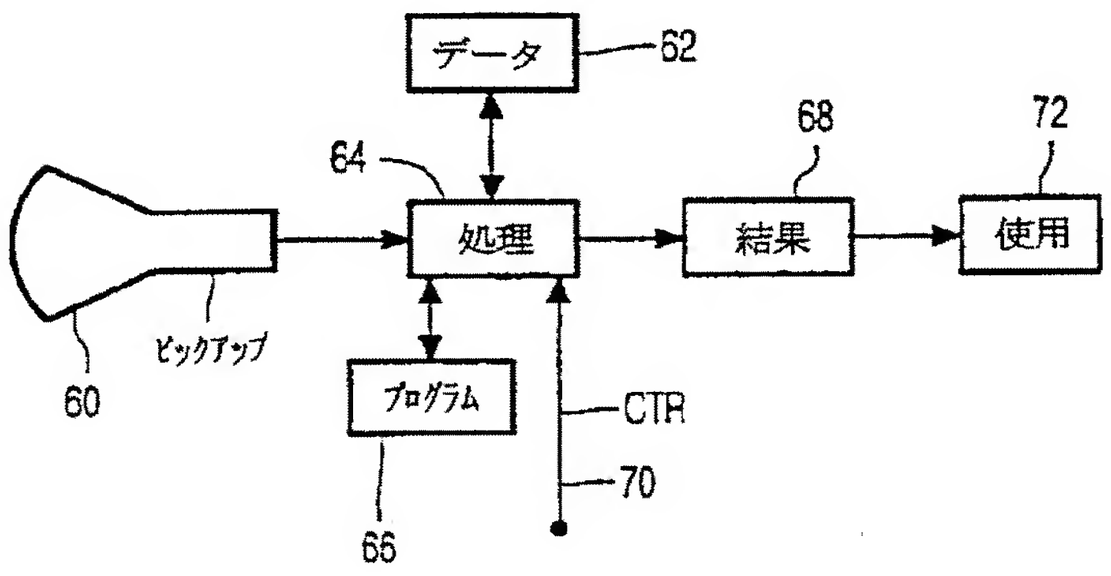


FIG. 2

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB 98/01990

A. CLASSIFICATION OF SUBJECT MATTER

IPC6: G10L 5/06
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC6: G10L, G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5680481 A (K. VENKATESH PRASAD ET AL), 21 October 1997 (21.10.97), page 2, line 25 - line 29; page 20, line 44 - line 46, abstract ---	1-5,9
A	WO 9733250 A1 (HEWLETT-PACKARD COMPANY), 12 Sept 1997 (12.09.97), claim 1 ---	1-3,5,9

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 June 1999	03 -07- 1999
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Peder Gjervaldsaeter/MN Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT
Information on patent family members

01/06/99

International application No.
PCT/IB 98/01990

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5680481 A	21/10/97	DE 4435272 A,C	13/04/95
		DE 4317372 A,C	02/12/93
		JP 6043897 A	18/02/94
		US 5586215 A	17/12/96
		US 5621858 A	15/04/97
		US 5771306 A	23/06/98
WO 9733250 A1	12/09/97	EP 0885427 A	23/12/98
		JP 9245124 A	19/09/97

Form PCT/ISA/210 (patent family annex) (July 1992)

フロントページの続き

(81)指定国 ＥＰ(ＡＴ，ＢＥ，ＣＨ，ＣＹ，
ＤＥ，ＤＫ，ＥＳ，ＦＩ，ＦＲ，ＧＢ，ＧＲ，ＩＥ，Ｉ
Ｔ，ＬＵ，ＭＣ，ＮＬ，ＰＴ，ＳＥ)，ＪＰ，ＵＳ